Exploratory Data Analysis

LIMITED PRELIMINARY EDITION

Volume II

by JOHN W. TUKEY

ADDISON-WESLEY PUBLISHING COMPANY
READING, MASSACHUSETTS
MENLO PARK, CALIFORNIA · LONDON · DON MILLS, ONTARIO

08.75

TABLE OF CONTENTS

Volume 2, Preliminary Edition

(For details, see tables of contents of individual chapters.)

Preface

Introduction to the preliminary edition
About the problems

16. Looking at elementary analyses

17. Elementary analysis is often a beginning

18. Expression or extension? (two uses of comparison values)

18x. Going further with our fit (still East Coast)

19. Analyzing more difficult tables

20. Blocks and unification (organized comparisons)

(No chapter 21 in this version)

22. Restrained fits (when numbers fade away)

23. Fitting can be confusing (posses and descriptions)
[essential background]

EXPLORATORY DATA ANALYSIS

VOLUME II

Limited Preliminary Edition

by

John W. Tukey

Princeton University
and
Bell Telephone Laboratories, Incorporated

The writing of this work, together with some of the research reported herein, was supported in part by the United States Government. Reproduction, translation, publication, use, and disposal in whole or in part, by or for the United States Government is permitted.

Copyright © 1970 by Addison-Wesley Publishing Company, Inc. Philippines copyright 1970 by Addison-Wesley Publishing Company, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. Published simultaneously in Canada.

TABLE OF CONTENTS

Volume 2, Preliminary Edition

(For details, see tables of contents of individual chapters.)

Preface

Introduction to the preliminary edition
About the problems

16. Looking at elementary analyses

17. Elementary analysis is often a beginning

18. Expression or extension? (two uses of comparison values)

18x. Going further with our fit (still East Coast)

19. Analyzing more difficult tables

20. Blocks and unification (organized comparisons)

(No chapter 21 in this version)

22. Restrained fits (when numbers fade away)

23. Fitting can be confusing (posses and descriptions)
[essential background]

PREFACE TO VOLUME 2

It is still well to understand what you can do before you learn how to measure how well you seem to have done it. This is even more true for two-way analyses, the subject of this volume, than for more elementary techniques. The other remarks in the Preface to volume 1 and in the Introduction to the Preliminary Edition apply with at least equal force to this volume -- please read them.

That introduction promised "a second volume". This is the first half of what was promised -- so there will have to be a "volume 3".

Techniques and concepts do build upon volume 1, but even the extreme short-cut version of that volume (see the back of its page I-3) should provide a basis for going ahead. The careful reader will note some discontinuities in details of format. I hope these reflect improvements. (I am confident of those which respond, at least in part, to Frederick Mosteller's suggestions.)

It is a pleasure to acknowledge support, guidance, cooperation and hard work. The Army Research Office (Durham), through a contract with Princeton University, and the Bell Telephone Laboratories have supported the writing financially. Charles P. Winsor taught the writer, during the 1940's, many things about data analysis that were not in the books. Although its formal beginning came after his death, it owes much to S. S. Wilks, whose leadership for statistics in Princeton made possible the gaining of the insights on which it is based. Careful reading, in various versions, by Dr. Leonard Steinberg and Mr. David Hoaglin have removed many errors, unclarities and awkwardnesses. (As always, the writer is solely responsible for those that remain.) Careful and skilled typing, mainly by Miss Elizabeth L. LaJeunesse (who later became Mrs. Elizabeth Dutka), and more recently by Mrs. Mary Bittrich, has been an essential.

It is important to read the Introduction to the Preliminary Edition which follows.

JOHN W. TUKEY

Princeton, New Jersey
October 16, 1970

INTRODUCTION TO THE PRELIMINARY EDITION

This is a slightly revised form of the third version of this account. Each version has been used in class in successive years. Each time enough has been learned to call for extensive revision and rearrangement. This was as true of the third trial as of the other two. This time, however, the signals are for a substantial reorientation rather than for the insertion of new material or the discussion of new techniques.

The preface has hinted at the separation of exploration from confirmation. What has not yet been recognized in this preliminary edition is the separation of breadth of possibility from seriousness of analysis. The First Edition, as seen dimly today, will try to bring the breadth of possibility into its first volume, making clear at the beginning some of the diversity of ways in which data can be explored. The second volume can then go back over these approaches in turn, explaining how each can be carried further.

We do have three obligations to our data, obligations we ought to meet one after another. We need to explore the data flexibly -- here flexibility can be enhanced by simplicity without exhaustiveness -- here is the place for Exploratory Data Analysis. We need to carry out the chosen analyses carefully and exhaustively -- especially if we are to use them

in a controversy (all publication is potentially part of a controversy, just as all use as evidence is potentially part of an adversary position -- more is the place for Mustering and Borrowing Strength. We need to assess their quality -- here is the place for Confirmatory Data Analysis.

The present version mixes first steps and detail together, treating each major topic from both points of view before passing on to the next. The teacher may wish to move closer to the structure planned for the First Edition by going quite selectively through both volumes of the preliminary edition. He cannot do this as well as he might; this version was not planned to let him do this -- but he can try. The guide to short-cutting that follows offers some help.

ABOUT THE PROBLEMS

The teacher needs to be careful about assigning problems. Not too many, please. They are likely to be longer than you think. The number supplied at the ends of the chapters is to accommodate diversity of interest, not to keep everyone busy.

Besides the length of our problems, both teacher and student need to realize that many problems do not have a single "right answer". There can be many ways to approach a body of data. Not all are equally good. For some bodies of data this may be clear, but for others we may not be able to

tell from a single body of data which approach is preferred. Even several bodies of data about very similar situations may not be enough to show which approach should be preferred. Accordingly, it will often be quite reasonable for different analysts to reach somewhat different analyses.

Yet more -- to unlock the analysis of a body of data, to find the good way or ways to approach it, may require a key, whose finding is a creative act. Not everyone can be expected to create the key to any one situation. And, to continue to paraphrase Barnum, no one can be expected to create a key to each situation he or she meets.

An eminent statistician once visited an eminent prep school teacher of physics. After the laboratory group was well into the experiment, the statistician hinted that it might be interesting to collect all the students results and display them together. "Oh", said the teacher, "but some of them will have gotten the wrong answer." The usually free-spoken statistician, whom long experience had taught to expect variation in everything, dared not trust himself to reply. To learn about data analysis it is right that each of us try many things that do not work -- that we tackle more problems than we make expert analyses of. We often learn less from an expertly done analysis than from one where we missed an opportunity -- at least until we were told about it. Each teacher needs to recognize this in grading and commenting on problems.

GUIDELINES FOR SHORTCUTTING AND DELAYING

Chapter	Various levels of shortcutting							
	Extreme[1]		Intermediate[2]		Mild[3]		None	
	Sections	*	Sections	*	Sections	*	Sections	*
16	A to G	29	A to G	29	A to G	29	A to G	29
17	------	--	(A to G	25)	(A to G	25)	A to G	25
18	------	--	(A to G	30)	(A to G	30)	A to G	30
18x	------	--	------	--	------	--	A to C	10
19	A to G	19	A to G	19	A to G	19	A to G	19
20	------	--	(A to G	27)	(A to G	27)	A to G	27
22	A to G	27	A to G	27	A to G	27	A to G	27
23	(A to D	19)	A to D	19	(A to F	24)	A to F	24
		{75/94}		{119/124}		{151/157}		191

* Pages of text (not exhibits). Note that sections H and P, which would be covered, are omitted from the count.

(1) with or without A to D of chapter 23.

(2) with any one of chapters 17, 18, or 20.

(3) with any three of chapters 17, 18, 20, and 23.